Московский государственный технический университет им. Н.Э. Баумана

Кафедра «Системы обработки информации и управления»

Рубежный контроль №1

по дисциплине

«Методы машинного обучения»

Выполнил:

студент группы ИУ5И-21М

Ван Чжэн

Москва — 2024 г.

Вариант:№.16

Я выбрал набор данных-«продажа подержанных автомобилей».

Дополнительные требования по группам: для пары произвольных колонок данных построить график "Диаграмма рассеяния".

```
df  =  pd.DataFrame(data.head(5000))

#  绘制散点图
plt.figure(figsize=(10,  10))
plt.scatter(df['Mileage'],  df['Price'],  color='blue')
plt.title('Mileage  vs  Price')
plt.xlabel('Mileage')
plt.ylabel('Price')
plt.grid(True)
plt.show()
```

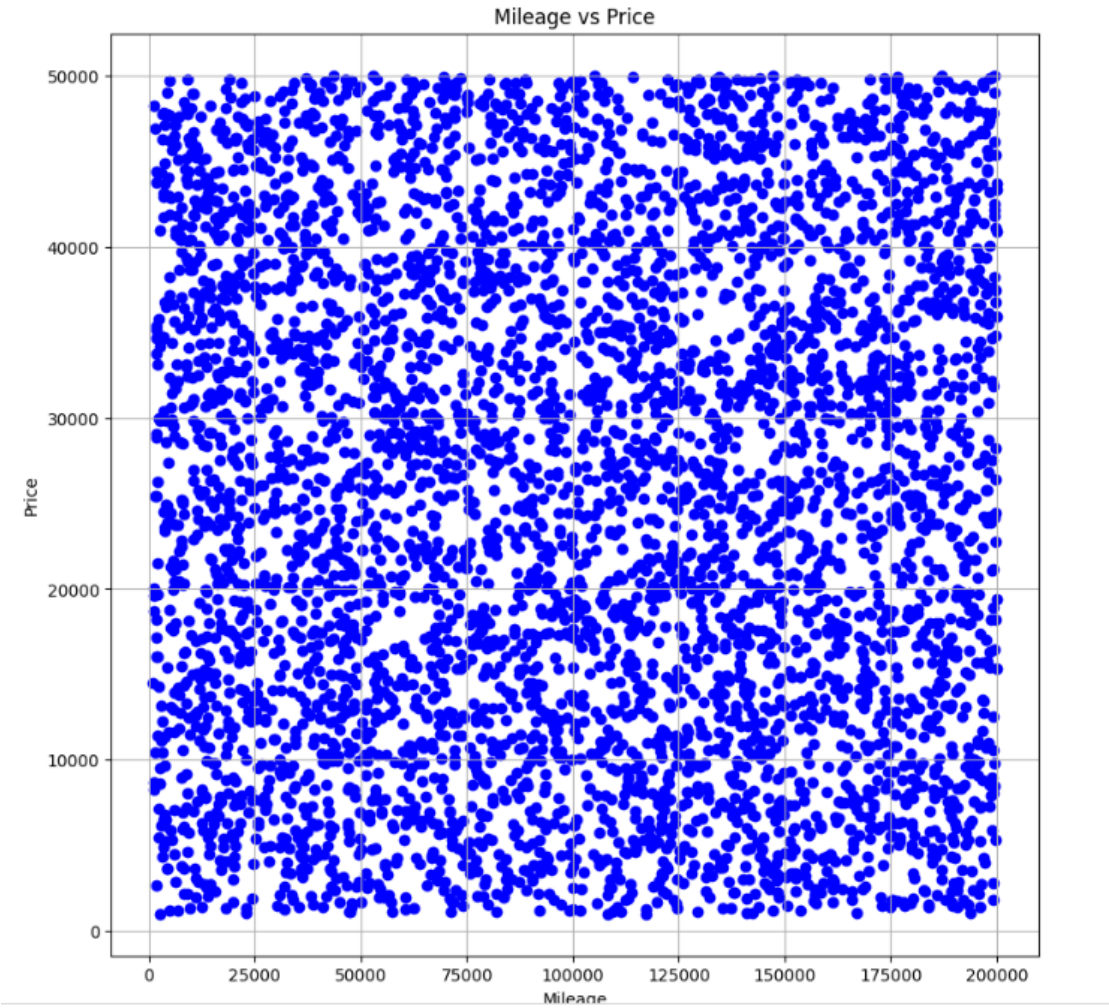


Рис1. - Диаграмма рассеяния

Задача1: №.16.

Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием преобразования Бокса-Кокса (Box-Cox transformation).

```
def diagnostic_plots(df, variable):
        plt.figure(figsize=(15,6))
        # Гистограмма
        plt.subplot(1, 2, 1)
        df[variable].hist(bins=30)
        ## Q-Q plot
        plt.subplot(1, 2, 2)
        stats.probplot(df[variable], dist="norm", plot=plt)
        plt.show()
```

```
[8]
    data['Price boxcox'], param = stats.boxcox(data['Price'])
    print('Оптимальное значение λ = {}'.format(param))
    diagnostic_plots(data, 'Price boxcox')
```

```
Оптимальное значение λ = 0.7289255862318366
```

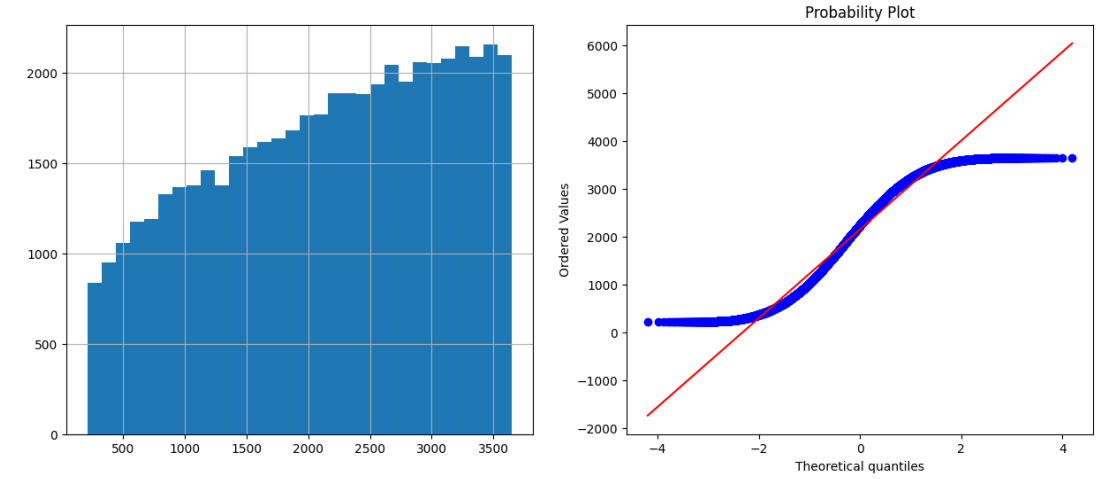
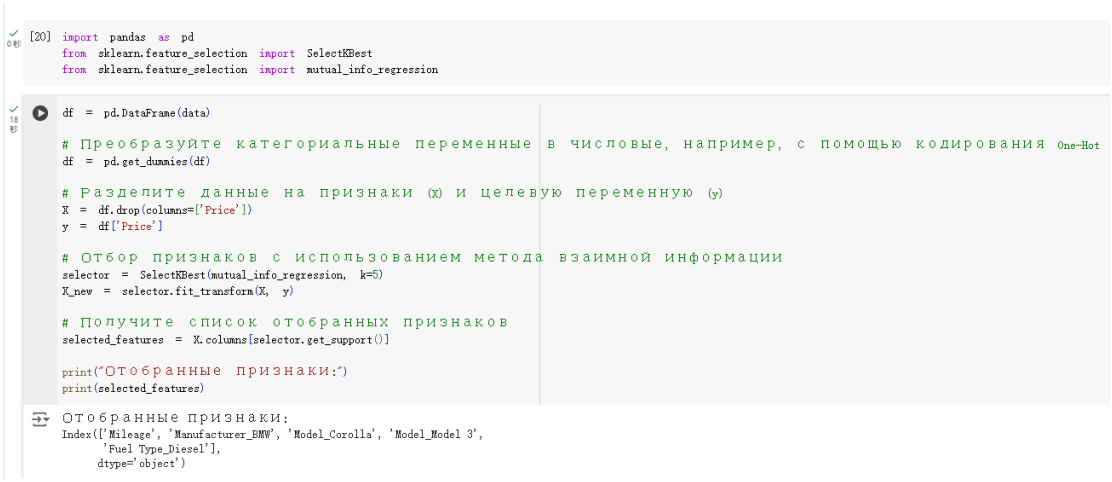


Рис2. - проведите нормализацию для «Price»

Задача2: №.36.

Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectKBest для 5 лучших признаков, и метод, основанный на взаимной информации.

Мы сначала создаем DataFrame из предоставленных данных, затем преобразуем категориальные переменные в числовые с помощью кодирования One-Hot. После этого мы разделяем данные на матрицу признаков (X) и целевую переменную (y). Далее мы используем класс SelectKBest с методом взаимной информации (mutual_info_regression) для отбора 5 лучших признаков. Наконец, мы выводим список отобранных признаков.



```
[20] import pandas as pd
     from sklearn.feature_selection import SelectKBest
     from sklearn.feature_selection import mutual_info_regression
```

```
df = pd.DataFrame(data)

# Преобразуйте категориальные переменные в числовые, например, с помощью кодирования One-Hot
df = pd.get_dummies(df)

# Разделите данные на признаки (X) и целевую переменную (y)
X = df.drop(columns=['Price'])
y = df['Price']

# Отбор признаков с использованием метода взаимной информации
selector = SelectKBest(mutual_info_regression, k=5)
X_new = selector.fit_transform(X, y)

# Получите список отобранных признаков
selected_features = X.columns[selector.get_support()]

print("Отобранные признаки:")
print(selected_features)
```

```
Отобранные признаки:
Index(['Mileage', 'Manufacturer_BMW', 'Model_Corolla', 'Model_Model 3',
       'Fuel Type_Diesel'],
      dtype='object')
```

Рис3. - проведите процедуру отбора признаков